DOCUMENT RESUME

ED 346 132 TM 018 427

AUTHOR Kahl, Stuart R.
TITLE Alternative Assessment in Mathematics: Insights from Massachusetts, Maine, Vermont and Kentucky.
PUB DATE Apr 92
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Educational Assessment; Elementary Secondary Education; *Mathematics Achievement; Mathematics Tests; *Portfolios (Background Materials); Program Evaluation; *State Programs; Student Evaluation; *Testing Programs
IDENTIFIERS *Alternatives to Standardized Testing; Kentucky Instructional Resource Information System; Maine Educational Assessment; Massachusetts Educational Assessment Program; Open Ended Questions; *Performance Based Evaluation; Vermont Mathematics Portfolio Assessment Program

ABSTRACT

Statewide assessment programs in mathematics that have led the way in the development and implementation of new alternative forms of assessment are described and compared. Assessments reviewed are: (1) the Massachusetts Educational Assessment Program (MEAP); (2) the Maine Educational Assessment (MEA); (3) Vermont's Mathematics Portfolio Assessment Program (VMPAP); and (4) the Kentucky Instructional Resource Information System (KIRIS). The four statewide assessment programs have a great deal in common. When the earlier ones (the MEA and MEAP) began, they were innovative, using matrix sampling to administer large numbers of items efficiently. Maine made significant use of open-ended questions. The two newer programs (the VMPAP and KIRIS) took a pioneer role in large-scale alternative assessment with the portfolio approach of Vermont and the portfolios and performance assessments of Kentucky. Some of the lessons to be drawn from these assessments are described. Recognizing the power of testing to influence instruction and curriculum, these states are using assessment as a vehicle for change while refining techniques for large-scale use. Fifteen pages of attachments provide sample questions and some of the forms used to administer the assessments. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
STUART R. KAHL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346132

ALTERNATIVE ASSESSMENT IN MATHEMATICS:

INSIGHTS FROM MASSACHUSETTS, MAINE, VERMONT AND KENTUCKY

Stuart R. Kahl
Advanced Systems in Measurement
and Evaluation, Inc.
171 Watson Road
Dover, NH 03820

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992

TM018427

BACKGROUND

Current trends in the assessment of mathematics are best described in a larger context. The education community generally is experiencing a mini-revolution in the area of assessment. Fueled by dissatisfaction with the quality of American education and with the capacities of traditional standardized achievement tests to meet the various needs of educators and consumers of education, this revolution has led to the development and implementation of assessment programs employing alternative forms of assessment. While still addressing the need to compare students or groups of students, new assessment approaches have been designed to do a better job of determining what students can actually do. Furthermore, recognizing the power of assessment instruments and practices to influence curriculum and instruction, reformers have tried not only to make assessment practices consistent with good instruction, but also to make instructional activities and assessment activities one and the same. This goal has been the greatest challenge for large-scale assessment programs because of demands of accountability and logistical concerns associated with large numbers of test takers. Nevertheless, many state testing programs have led the way in the development and implementation of the new, alternative forms of assessment.

Innovative large-scale testing programs are particularly interesting in light of their status relative to several controversial issues in testing. Some of those issues and related terminology are discussed below.

Direct vs. Indirect Assessment

Traditional standardized achievement tests have been used effectively to compare students for a long time. Years ago, nobody claimed that standardized, multiple-choice achievement tests measured directly the kinds of competencies students should have been developing in order to function successfully in life. Instead they were proxies or indirect measures of more authentic, "real-world" performances. Again, for a limited number of purposes requiring the comparison of individual students, they worked. Today teachers continue to find that their stronger students score higher on such instruments than their weaker students.

However, increasing concern for school accountability and the lack of alternative forms of assessment, caused these tests to be used for purposes for which they were not optimally designed — instructional program evaluation and curriculum assessment. School curricula in all subjects were soon defined in terms of specific concepts and isolated skills — the kinds easily measured by the individual multiple-choice questions appearing in the standardized achievement tests. And the easiest way to raise test scores was to focus instruction on these specific concepts and skills in isolation. In effect, educators turned the traditional indirect measures of desirable outcomes into direct measures of less desirable outcomes, not by changing the tests, but rather by atomizing curricula and instruction, making specific pieces of knowledge and isolated skills the goals of mastery- oriented programs.

BEST COPY AVAILABLE

Low Stakes vs. High Stakes Testing

Any number of factors might be credited with leading to the abandoning of the mastery-oriented instruction that has so dominated American public education the last few decades. Those factors are not a major topic of this paper. However, one important development leading to the disfavor into which tests consistent with such instruction have fallen is of interest here — namely, the Lake Wobegon Effect. This phenomenon is essentially the sense of complacency that arises as a result of test scores suggesting a level of competence for students that the students have really not attained. The standardized test scores can be inflated or otherwise misleading for several reasons ranging from unethical practices of school staffs to the more ethical, but still harmful, narrowing of the curriculum to match the tests. Such behaviors are facilitated by the lack of test security (the tests remain in the schools for years) and are practiced because of the pressures on teachers and administrators to raise test scores.

These pressures are associated with high stakes testing programs, generally accountability testing programs producing school or district results that are highly visible and can be used, not necessarily appropriately, to evaluate the educators. Some educators or testing critics would argue that tests should never be high stakes tests. However, the advocates of high stakes testing argue that low stakes tests produce underused results and are low in impact. That some tests seem to have negative impacts on curriculum and instruction is seen as an indication of the potential for more desirable tests to have positive impacts.

On Demand vs. Integrated Assessment

Traditional testing is generally conducted on an on-demand basis. That is, the tests are administered in tightly controlled situations, usually with strict time constraints. Test security is important to maintain before, during, and after such testing. Some educators argue that on-demand testing is not "authentic," that when performing "real-life" tasks, people have time to reflect, revise, seek advice, etc. Critics of on-demand testing, favor integrated assessment instead. Such assessment consists of activities that are part of regular instruction over time — not as traditional testing is interspersed with instruction throughout the year, but rather as student work other than on formal tests is evaluated by teachers every day. Portfolio assessment as implemented in some states seems to be the best example of integrated assessment on a large scale. Student work on "longer-term" activities is collected throughout the school year. Portfolio "entries" are treated much like students' compositions in a class in which the "writing process" is employed. Students produce work over a period of time, it may be discussed with other students, feedback is given, revising is done. Students and teachers jointly decide on "best pieces" to be included in the portfolio submitted in conjunction with the statewide testing program.

It is probably through integrated assessments that a state department can have the most desirable impacts on curriculum and instruction. However, there are concerns from the accountability perspective that the lack of control over many factors is problematic. At some point, it is important to know what a student or students can do on their own. Furthermore, there are many who feel that much of the work people produce after formal schooling has ended, they produce on demand. It is important to note that innovative testing formats can still be on demand. For example,

the administration of performance tasks can be accomplished under closely controlled, timed, secure conditions. Traditional direct writing assessments are examples.

Internal vs. External Assessment

When teachers are intimately involved in the selection of tasks or task options for their students and also involved in the evaluation of their students' work on those tasks, then internal assessment is taking place. When a state department produces testing materials, ships them to schools for teachers to administer, gets them back from the schools, and has the students' work scored (even if by teachers at state scoring sessions), then external assessment is taking place. As more and more emphasis is being placed on performance-based assessment, aspects of internal assessment are being relied upon more, even for larger scale testing programs. This situation has come about largely for logistical and cost-related reasons. Of course, if school change is more likely to occur when the stakeholders are involved directly in the process, then the incorporation of internal components in a state's assessment program can be an effective means of encouraging the kinds or curriculum reforms and instructional improvement professional education groups are recommending.

Individual vs. Group Results

In the past decade, educators have become increasingly aware that assessment instruments optimally designed for one purpose may differ considerably from assessment instruments designed for a different purpose. Of course, for program or curriculum evaluation for which individual student results need not be produced, the efficiencies associated with matrix sampling have been extremely beneficial to large statewide testing programs. (Matrix sampling involves the development of a very long test in a subject, that test then being divided into many different, probably nonequivalent forms, and administered one form per student. Extremely reliable school level results can be produced even through each student may respond to just a small number of questions.) Besides the potential to use matrix sampling effectively, instruments designed to produce only group results could differ from individual tests in many other ways relating to domain coverage, item characteristics, etc.

With state departments of education taking on greater responsibility for assessing students and school programs, more efficient techniques are continually being sought so that assessment programs of tremendous scope can be feasible both from a cost and a management perspective. Decisions to produce only group results lead to major savings. When individual student results are required, states are designing programs with different components, recognizing that instruments yielding individual scores would not provide the kind of information needed for program evaluation. As the desire for performance-based assessments increases, so does interest in and the necessity of efficient practices. Again, matrix sampling can be helpful, but so can assessment components that are integrated or internal assessments. Shifting some responsibilities to the local level may be the only way extensively performance-based assessments in some states are feasible. This is a rationale behind the design of several innovative statewide testing programs. Further justification for the added burden on school personnel is that it should not be an added burden at all since the assessment activities are what students should be engaging in as part of their regular instruction. A statewide program that is a combination of an external, on-demand component (as performance-based as

possible) producing group results, and an internal, integrated component producing individual results has many advantages.

The statewide assessment programs described in later sections of this paper differ with respect to their status relative to the issues discussed above. However, all of these programs have been evolving over time, and changes that have been made in them or that are being considered are closely tied to these issues.

Movements in Mathematics Education

The New Math, Back to the Basics, The National Council of Teachers of Mathematics' (NCTM's) Agenda for Action, NCTM's Curriculum for the 80s — every decade has seen a new effort by mathematics educators to reform school mathematics. That these efforts had any significant positive impact is highly suspect. However, it does not appear that a lack of impact will be a problem for NCTM in implementing its recent *Curriculum and Evaluation Standards for School Mathematics*. The NCTM Standards spell out in great detail what mathematics curricula and instruction should be like. NCTM encourages a holistic view of mathematics emphasizing understanding, not rote learning; applications, not abstractions; problem solving, not drill; thinking, not recall. While a lot of school programs need to be changed to be more consistent with this vision, intensive effort is being made at all levels to bring about the necessary changes.

One reason the NCTM Standards are off to a good start is simply timing. Dissatisfaction with American schools and their showing in the international arena, dissatisfaction with traditional tests which many believe perpetuate poor instruction, and increasing interest in developing higher order thinking skills have triggered many efforts to reform schools in general as well as mathematics programs specifically. Recent projects intending to develop ambitious national standards and assessment methodologies consistent with them have given the NCTM Standards a stamp of approval, commending NCTM for accomplishing an initial step toward reforms that remains to be done in other disciplines. The proceedings at the National Summit on Mathematics Assessment sponsored by the Mathematical Sciences Education Board last spring constitute one such endorsement.

It is not surprising that state departments of education are changing the way they assess the mathematical competencies of students and the effectiveness of instructional programs in mathematics. All of the statewide testing programs discussed below have

- avoided describing the content of test instruments in terms of specific skills or narrow behavioral objectives,
- increased their use of non-multiple-choice measures,
- increased their emphasis on problem solving and decreased their emphasis on skills in isolation (e.g., computation),
- implemented or made plans to implement performance or portfolio assessment.

The Massachusetts Educational Assessment Program (MEAP)

Massachusetts' program began in 1985-86 when all students in grades 3, 7, and 11 were tested in reading, mathematics, and science. Testing takes place every other year in that program which tests 180,000 students every round. Starting with the 1987-88 school year, the grades tested were 4, 8, and 12; and social studies was added to the subjects assessed. Matrix sampling is employed to ensure that several hundred multiple-choice questions in every subject are administered in every school. Only school and statewide results are produced. Because of the broad domain coverage provided through matrix sampling, school scores are reported for many subcategories of mathematics. Tests in successive cycles are statistically linked to facilitate the monitoring of changes in performance.

For the first three rounds of MEAP, open-ended questions were administered on a sampling basis, and the data from them were used only to produce statewide item results discussed in narrative, interpretive reports. Attachments 1a, 1b, and 1c illustrate the kinds of open-ended questions asked in the last three rounds of the program. These are "extended open-ended questions," not just short answer questions. Generally the MEAP open-ended questions stress competencies NCTM values greatly — communication, reasoning, and problem solving. Responses were scored analytically in that each response was assigned to one of a large number of well defined categories.

This year, every student has responded to at least one of the matrix sampled open-ended mathematics questions. The results on the open-ended questions will figure prominently in the school level results in mathematics. Each student's response is being scored holistically on a scale from 1 to 5. Every open-ended question has its own scoring guide developed to be consistent with a general scoring rubric describing "1" responses as completely incorrect or irrelevant and describing "5" responses as showing complete understanding of the problem, using appropriate methods of solution, demonstrating clear reasoning and communication, containing no significant errors, and providing effective examples where appropriate. This general rubric was used as the basis for developing tailored scoring guides for open-ended questions in all four subjects assessed.

In 1989, statewide samples of students in grades 4 and 8 participated in performance testing in mathematics. Trained administrators (teachers in Massachusetts) administered performance tasks to pairs of students. The tasks involved the use of mathematical tools and manipulatives. The administrator's script was also the place where he or she recorded the students' actions and responses. Attachments 2a, 2b and 2c describe three of the performance tasks administered in Massachusetts.

Current plans in Massachusetts call for MEAP testing to rely exclusively on open-ended questions eventually. Also, beginning this year, open-ended questions and the descriptions of responses at different levels in the scoring guides will be used to define proficiency levels in every subject area. The emphasis in reporting will shift from average scaled scores to percentages of students at the different proficiency levels in each school and statewide .

The Maine Educational Assessment (MEA)

The MEA is in its seventh year of operation. Annually it tests all 4th, 8th, and 11th grade students in reading, writing, mathematics, science, social studies, and the humanities. The program is really two programs in one in that: 1) a common set of questions in reading, mathematics, and writing is administered to all 15,000 students at a grade level, yielding individual students results in those areas, and 2) a larger number of items in reading, mathematics, science, social studies, and humanities are administered in a school through matrix sampling to produce reliable school and subgroup results. In the past the program relied primarily on multiple-choice questions with some open-ended questions in reading and mathematics and a writing prompt eliciting writing samples (compositions) from students.

The MEA, like Massachusetts's program, has not been driven by a lengthy set of specific objectives. Also, there has been some shifting of emphasis toward higher order skills, although the major emphasis from the beginning was on such skills. One important change has been in the mathematics reporting category of "Procedural Knowledge." Originally, this category included skill-level questions (e.g., computations) predominantly. In recent years, routine story problems have constituted the better part of this category of items.

The first five years of the program, ten of the fifty "common" mathematics questions at a grade level were open-ended questions. While the questions were non-trivial, the space allowed for responses was somewhat limited, all ten responses on one page. The responses were scored analytically, and each open-ended question "counted" the same as a multiple-choice question. In year six, the responses to open-ended questions were scored holistically on a scale from 0 to 4, and each open-ended question then "counted" 4 times as much as a multiple-choice question insofar as individual results were concerned.

This past year, the use of open-ended questions was expanded through the use of matrix sampling. Each student responded to five common open-ended mathematics questions and an additional question unique to his/her test form. There were ten forms. Thus fifteen open-ended questions were administered in every school, they were scored holistically, and they accounted for approximately 30 percent of each school's score in mathematics. The space available for responses to the common open-ended questions was doubled (five responses per page), and a half a page was available for each student's response to the matrix sampled question. Attachments 3a and 3b are a set of common open-ended mathematics questions and the corresponding response page.

The MEA has found ways of having students use manipulative and mathematical tools to respond to some mathematics questions. Some preliminary work, especially at grade 4, was required to prepare the manipulatives for use since they were provided to each student in the form of a separate perforated sheet or a sheet requiring cutting along dotted lines. One-inch square counters and tangram pieces were provided this way. Also, some questions required the use of a paper ruler provided on the separate sheet. Attachments 3c and 3d are sample manipulatives that have been used in the MEA.

Two years ago, a small study requiring the use of calculators was conducted in Maine. Also information was gathered from school staffs on the availability of calculators to students. In 1991-92, it was required that a calculator be made available to every eighth and eleventh grade student during two of three mathematics testing sessions. (The third session was a very short session during which only ten noncalculator, multiple-choice questions were administered.) The calculators were NOT provided by the state.

Plans for the MEA call for increased emphasis on open-ended questions in the immediate future. Also, a small pilot study involving mathematics portfolio assessment is to be conducted this coming year. Tentative plans for the following year include portfolio assessment involving a statewide sample of schools, with the possibility of full-scale implementation of portfolio assessment statewide the year after that.

Vermont's Mathematics Portfolio Assessment Program

After two years of planning, Vermont conducted a mathematics portfolio assessment pilot study at grades 4 and 8 in 1990-91. The primary purpose of this study was simply to gather portfolios so that the proposed scoring criteria could be tried out and refined. In 1991-92, every grade 4 and 8 Vermont student maintained a mathematics portfolio. This May, at regional scoring sessions, a statewide sample of the portfolios will be scored. Next year, when the program is fully implemented, all the portfolios will be scored by the students' own teachers. A moderation system will by used to assure that the teachers' scoring is "on target" and corrective action will be taken if necessary. The moderation process involves cluster scoring sessions. Several schools belong to a cluster, and each cluster has a mathematics portfolio cluster leader who runs the scoring session. Each teacher of grade 4 or 8 students in the cluster, brings a sample of his or her students' portfolios to the session. (The sample is determined by the state department of education.) The teachers are trained, and then they rescore the portfolios (not their own) at the scoring session. If there are acceptable levels of score agreement with respect to the portfolios scored by a particular teacher, then the scores the teacher originally gave the portfolios in his/her school are allowed to stand. Otherwise, the cluster leader must coordinate or conduct retraining of the teacher and rescoring of the portfolios of his/her students. With the full implementation of the system, district level scores will be produced in addition to the individual portfolio scores obviously produced.

Attachments 4a, 4b, and 4c describe the required portfolio contents and the scoring criteria. The focus is on problem solving and communication. The students maintain working portfolios throughout the school year. Portfolio entries are students' work on various kinds of mathematical problems or projects (e.g., puzzles, investigations, applications). The production of a portfolio entry is much like the production of a writing sample in a classroom in which the "writing process" is employed in that a student can revise his/her work, obtain feedback, etc. Entries can be in any of a number of forms — e.g., written problem solutions, reports, videotapes, posters. For purposes of statewide assessment, a student and his or her teacher jointly identify the 5 to 7 entries to be submitted for scoring. Those entries must represent a range of entry types and involve a range of mathematical content and contexts.

The portfolio's are scored using four problem solving criteria and three communication criteria. (See Attachment 4c — the Mathematics Portfolio Profile Worksheet.) Thus, each portfolio receives seven ratings. Tallies are recorded on certain fields of the worksheet to monitor the breadth of coverage of content and other task characteristics.

The description of Vermont portfolios thus far does not communicate the scope of the task of implementing portfolio assessment statewide. Perhaps the greatest effort is expended in teacher training and support. This is because the maintaining of portfolios as prescribed by Vermont's program requires a great departure from the normal teaching practices of most teachers. Large numbers of well attended workshops for teachers have been held throughout the state, including eight five-day institutes last summer. In addition to other workshops during the course of the year, cluster leaders give a great deal of support to teachers, providing them with materials and training. The major topics of training include: 1) the characteristics of good, rich performance tasks generating worthwhile portfolio entries; 2) effective portfolio management; 3) instruction consistent with the goals of the portfolio assessment program; and 4) the scoring of portfolios using state criteria.

As stated earlier, the more immediate plans for the portfolio assessment program in Vermont involve the local scoring of all students' portfolios, with that scoring moderated at the cluster level. The reporting of statewide, regional, district, and individual results is intended. Long range plans call for portfolio assessments at other grades and in other subjects.

Vermont also has a uniform mathematics assessment currently being administered to statewide samples of grade 4 and 8 students on an on-demand basis. Although smaller in scope, this assessment resembles the Massachusetts assessment employing multiple-choice and open-ended questions and matrix sampling. Several research questions should be answered after data are generated from the various assessment efforts.

The Kentucky Instructional Resource Information System (KIRIS)

The legislature in Kentucky mandated what is currently the most ambitious statewide assessment program in the country. In one way or another the program involves all modes of assessment, all grade levels of students, and all school subjects. In 1991-92, reading, writing, mathematics, science and social studies were assessed. In later years, additional areas will be assessed — e.g., arts and humanities, practical living skills, vocational studies. Test development in all areas is guided by six learning goals and several broadly defined valued outcomes determined by task forces working for Kentucky's Council on School Performance Standards.

The high stakes portion of the program is the accountability testing taking place at grades 4, 8, and 12. One component of that testing is administration of "transitional tests" consisting of multiple-choice and open-ended questions in reading, mathematics, science, and social studies and an on-demand writing sample. The transitional tests make use of common and matrix sampled questions. As in Maine, all students (50,000 at each of the three grade levels) answer the common questions yielding individual student results. The matrix sampled questions improve the coverage of domains for the production of meaningful program level results. At grades 8 and 12, the mathematics test required that all students have calculators. The assessment of writing portfolios is

also being accomplished in 1991-92. A moderation system similar to the one described for Vermont is being used. In 1992-93, mathematics portfolios will be assessed, and in subsequent years additional areas will be assessed via portfolios.

Also in 1991-92, on-demand performance events (tasks) in mathematics, science, and social studies were administered to students in all schools in the accountability testing grades. At each grade, four performance events were used in each of the three subjects. A trained facilitator/administrator visited each school in the state for a full day. In most schools, all grade 4 and 8 students participated. In any school with more than 150 students in a grade, a random sample of 90 students participated. This was the case for grade 12 in most high schools.

In each school, a subset of the twelve grade-appropriate tasks were administered in each of several hour-long sessions. The students worked at stations set up by the facilitators with the help of teachers. The stations were equipped with tools, manipulatives, resources and other materials required by the tasks. Most of the tasks were designed to involve both group and individual work. For example, some preliminary investigations may have been performed and discussed by a group of students, and then the students worked individually on an application of what was learned during the group work. Attachments 5a and 5b are grade 12 mathematics performance tasks recently administered. Each student worked on only one task (except in very small schools), and each student turned in a scorable product or products based on his/her work on the task. The products always included responses on a student direction/response form, but sometimes included other products such as a poster. Since the performance testing used matrix sampling (students took different tasks), the results of this component are only being used for school results. In the future, in addition to performance testing being done in more subject areas, greater use of technology (e.g., video monitors and cameras, computers) is planned for the performance events, both for stimuli and student products.

As indicated previously, mathematics portfolios will be assessed statewide in 1992-93. While materials related to this component are not yet releasable, it is safe to say that the Kentucky portfolio assessment in mathematics will bear some similarity to that in Vermont in terms of portfolio content and management, the need for teacher training, and the moderation process. Current plans for scoring, however, call for a single, holistic score to be assigned to each portfolio. Also, instead of focusing on individual pieces on the scoring worksheet, teachers will probably produce preliminary ratings for whole portfolios on problem solving, reasoning, communication, and integration of ideas (content).

As suggested previously, Kentucky's intent is to expand the use of performance and portfolio assessments. At the same time, tentative plans include the reductions in the use of the more traditional "transitional" testing (multiple-choice/open-ended). In the non-accountability grades, students are also being tested using instruments mirroring those used in grades 4, 8, and 12. Again, this "scrimmage" testing is to be expanded in scope over time. Results from the scrimmage testing are to be reported in much the same way as accountability results are, except high stakes are not being attached to them. The cognitive data from the accountability grades are to be merged with other school effectiveness indicators to produce an overall success indicator for every school. Then over time schools are to progress toward target figures for indicators determined uniquely for each school. Distinguished educators will investigate schools not progressing as they should.

Insights and Conclusions

The four statewide assessment programs described in the previous sections have a great deal in common. At the time the earlier ones (Maine and Massachusetts) began, they were quite innovative. The rejected the specific skills orientation that was so prevalent in statewide testing at the time. They both used matrix sampling to efficiently administer large numbers of items, thereby allowing the reporting of detailed school-level results. One (Maine) made significant use of open-ended questions in reading and mathematics and performance testing in writing. Both programs, however, have changed over the years. They have chosen to diminish their reliance on multiple-choice questions and move toward more performance-based approaches.

The two newer programs (Vermont and Kentucky), still in their infancy, moved directly to a pioneer role in large-scale alternative assessment, Vermont with its portfolios and Kentucky with both portfolios and performance assessments on a grand scale. Educators have been anxiously awaiting information from these programs about the effectiveness of their nontraditional assessment methodologies. Now information is available on how the assessments were conducted and on what procedures seemed to work well and which ones did not. Unfortunately, as of this writing, it will be a few more months before a great deal of information on the technical quality of the performance and portfolio assessment methods becomes available. Nevertheless, the programs reviewed above provide us with some useful insights.

- Not only should state testing programs model desirable teaching practices, they should also model feasible ones. The Massachusetts performance testing used some interesting mathematics tasks which were administered to pairs of students by trained administrators in half-hour sessions. Although the teachers in Massachusetts appreciated the quality of the tasks, they believed the approach was not reasonable to emulate in a classroom with 25 students and one teacher. The Kentucky performance testing, however, modeled a system by which whole classes of students could be meaningfully occupied at one time. The approach is not unlike traditional high school chemistry labs except that students in Kentucky were involved in different tasks during a testing session. The requirement of scorable products meant that the administrator did not have to observe each student's every move.

- In some states (other than those discussed herein), advisory groups have strongly suggested that calculators be required equipment for testing. Yet the state departments have resisted the move believing that an inequity might be created since disadvantaged students may have more limited access to calculators and therefore the state might have to provide calculators for all students. Yet in Maine, given adequate warning, school officials were quite accepting of the requirement of calculators provided by the schools or the students themselves. Kentucky, which also required that calculators be made available, took an interesting stance regarding the inequity issue: if a group of students has limited or no access to calculators, then the inequity is in their instructional program, not the testing. If calculators are a recommended tool for instruction, and they are, then a school whose students do not have them <u>should</u> perform poorer than other schools on the test, all else being equal. This same

stance was taken in Kentucky with respect to equipment the schools were to provide for performance testing.

- Most of the programs described in this paper had predecessors that had low stakes. The general feeling within the states is that those programs had little or no impact. The impact of high stakes testing on curriculum and instruction has been well documented. Since what is tested is what is taught, the four programs are moving in the right direction by focusing on problem solving, reasoning, communication, and integration of knowledge through "bigger" tasks, particularly if NCTM's vision is a goal.

- The portfolio assessment approach being used in Vermont and Kentucky leaves a great deal up to the students and teachers. This lack of control may lead to inappropriate discrepancies between results from portfolios and from other modes of assessment. Until proven otherwise, it may be advisable that external on-demand testing be continued in conjunction with internally controlled integrated assessments. Kentucky has its on-demand transitional testing and performance testing; Vermont has its uniform mathematics assessment. Actually, the information on interrelationships among testing modes will be useful.

- If performance-based methods are to be employed economically on a large scale, significant cost-saving and time-saving measures must be taken. Some efficiencies are provided by the matrix sampling of tasks as in Kentucky's performance testing and by a portfolio assessment system in which teachers play a vital role, even in scoring. A desirable model for an efficient state assessment program might include: 1) an external, on-demand test as performance-based as possible through the use of matrix sampled open-ended questions (to produce school results); and 2) a portfolio assessment with local scoring and moderation as in Vermont and Kentucky (to produce student level results).

- Many people assume that portfolio assessment will force a positive change in instruction. In Vermont's pilot testing, some portfolios contained nonscorable entries such as drill sheets. Of course, if the results had counted, perhaps more appropriate entries would have been submitted instead. However, even then, a teacher could assign a rich task to students just five times during a year; then use undesirable teaching practices the rest of the year, still satisfying the portfolio requirements. Portfolios alone will not solve the problems of education. Teacher training and support is essential.

- If there is a data quality problem associated with performance or portfolio assessments, it is more likely a problem of generalizability of results due to limited domain coverage than it is a problem of scoring accuracy. Scores on writing samples elicited by different prompts are correlated approximately 0.5. It may be that nine or ten writing samples (or extended mathematics performance tasks) would be required to produce an acceptable level of generalizability. Where portfolios fall on the generalizability continuum is yet unknown.

- Responses to open-ended questions can be scored accurately and efficiently. Scored holistically on a scale (e.g., from 1 to 5) and therefore discriminating among students at several points along an ability continuum, one "extended" open-ended question can be worth two or three multiple-choice questions in terms of its contribution to reliability via internal consistency. Such a question requires one-half to one page for workspace and response and approximately eight to ten minutes of testing time. Rates of scoring the responses to open-ended questions can be quite variable and susceptible to change. Hints for speed and efficiency of scoring include:

  - use single readings where appropriate (e.g., if only school results are being produced or for students scoring far from a significant cut score);
  - use scorers with content expertise if possible;
  - keep the number of different questions scored at one time small;
  - monitor scoring rates and provide ambitious target rates.

As indicated in an earlier section, many attempts to reform mathematics curriculum and instruction have been made over the years. However, they did not coincide with a promising reform effort in testing that is totally consistent with the vision NCTM has of where mathematics instruction should go. Recognizing the power of testing to influence curriculum and instruction, states such as Massachusetts, Maine, Vermont and Kentucky are making the most of this opportunity to use assessment as their vehicle to change curricula and instruction. At the same time, these and a few other states are refining techniques that will soon be employed on a much larger scale. A great many states are planning new performance-based assessment programs.

Grade 12

Student Name:
School Name:

MATHEMATICS QUESTIONS

4. In a recent survey, Americans were asked about ownership of firearms. The findings of the study were that

   - 25 percent of American families have at least one handgun;
   - 25 percent of American families have at least one rifle; and
   - 10 percent of American families have at least one automatic rifle.

   A reporter used the following headline on an article she wrote about the study:

   MAJORITY OF AMERICAN FAMILIES OWN FIREARMS

   Should the reporter's editor accept the headline as it is? Why or why not?

   _____

   _____

   _____

   _____

   _____

   _____

   _____

   _____

5. Roger says that raising the score on a high-scoring test paper would raise the class average on the test more than raising the score on a low-scoring paper by the same amount. Is Roger right or wrong? Use the space below to explain or prove your answer to someone who does not agree with you.

8. On John's tenth birthday, John's grandfather gave him $10. He gave John $20 on his eleventh birthday and $40 on John's twelfth birthday. Following this pattern, John's grandfather plans on giving John $70 on his thirteenth birthday, but John expects $80 from his grandfather on that day. John's sister says that both amounts could be correct.

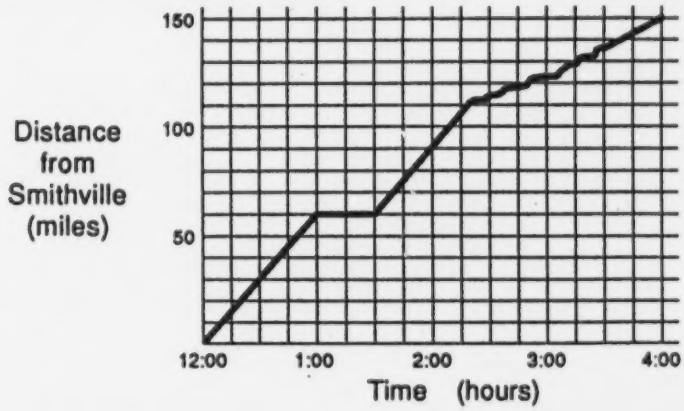
Who was right - John's grandfather, John, or John's sister? _____

Explain your reasoning. _____

9. Rhonda computed 5 X 496 in her head in just a few seconds. Explain how she probably computed this so quickly without paper and pencil.

PLEASE GO ON TO THE NEXT PAGE→

11. The Mitchells took a 4-hour car trip from Smithville to Guilford. The map and graph below help describe their trip.



Use the map and the graph to describe what the Mitchells were doing during each 1-hour interval. Tell as much as you can about the trip (e.g., kinds of roads, traffic jams, stops, etc.).

12:00 to 1:00 _____

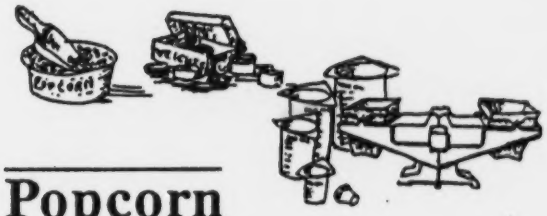
1:00 to 2:00 _____

2:00 to 3:00 _____

3:00 to 4:00 _____

This activity was adapted from an activity developed by Malcolm Swann, Shell Centre for Mathematical Education, University of Nottingham, 1985.

PLEASE GO ON TO THE NEXT PAGE →



Popcorn Estimation

Materials:

kernels of popcorn in original container

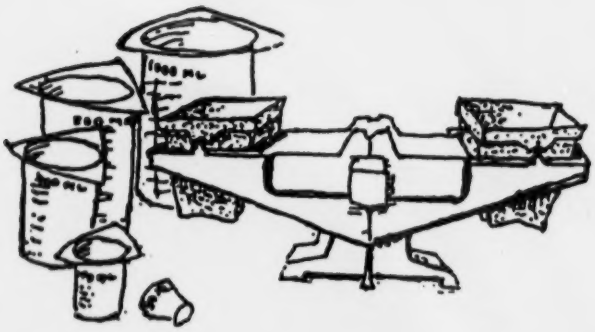
scoop

30 ml medicine cup

set of containers of different sizes marked in millilitres

balance with pans

a set of weights totalling 120 grams

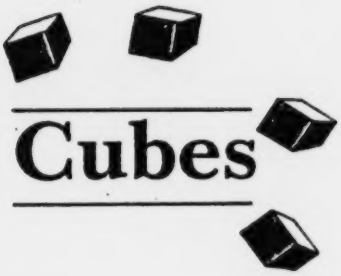


Description

Popcorn Estimation was the least structured and the most open-ended of the problems presented in this series. Students were told to use whatever they wanted from a set of materials to estimate the number of kernels in a container.

> "The only thing that you can't do is count all the kernels in the container. The closest estimate to the real number would be the best estimate for me."

The role of the administrator was limited to that of an observer and recorder. When the students arrived at their solution, the administrators checked the accuracy of their own observations by asking the students to describe their method. Students were then asked to suggest an alternative strategy for solution. If appropriate (i.e., if students were able to describe a second method and if there was time), they were asked to repeat the task using their proposed method. Twenty-eight percent of fourth graders and 66 percent of eighth graders did this.



Cubes

Problem 1

Students were asked to determine how many different rectangular solids could be made with 16 cubes. Administrators demonstrated a rectangular solid if necessary. They also gave the students a chart and suggested that they use it to keep track of the different shapes.

Materials: 16 wooden cubes

a chart with columns labelled height, length, width.

16 Cubes

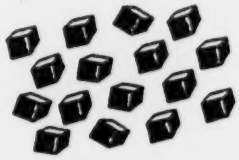
height	length	width

Problem 2

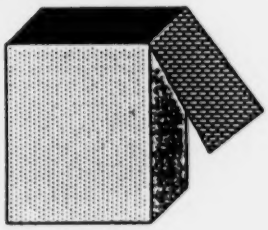
The students were given a container and asked to find out the maximum number of cubes that would fit.

Materials: 16 wooden cubes

container with a capacity of 96 cubes



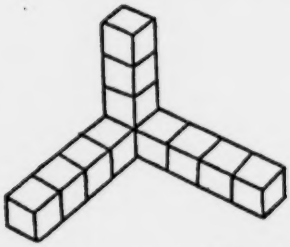
16 cubes



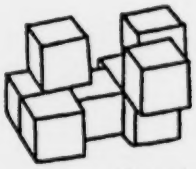
Container

Problems 3, 4, and 5

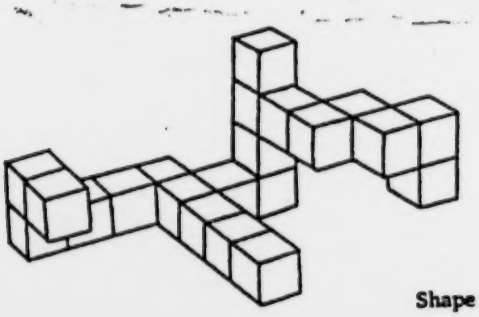
The administrator presented Shape A, B, and C in turn (Shape C was not presented to Grade Four), and asked "How many more cubes are needed to fill in this rectangular solid without making it any bigger?" The original 16 cubes were also available as a concrete aid for students.



Shape A



Shape C



Shape B

Math Town

Description

This task was adapted from an instructional unit designed by Judy White of the McCarthy Middle School, Chelmsford.

Materials:
- Town of Clear River map
- Set of cards with facilities and restrictions
- Ruler
- String
- Compass
- Transparent grid
- Paper and pencil

The administrator presented all the materials, except the cards, to the students and described various important features on the map such as the highway, the river, the park, and the town limit. (These features all had some significance in solving the problem.) The administrator then gave the students the cards that listed the facilities to be located and their restrictions. Students were told that they should work together and discuss their placements before committing them to paper. They were asked to consider other factors besides the ones listed on the cards in order to come up with the best location. The following problems were given:

Place a bicycle path within Clear River Park. It must be 5 1/2 miles long and continuous. It must start and finish at the boathouse.

Place a factory. There can be no houses within one mile. It must be within the town limits.

Place a regional school. It can be no more than 1 1/2 miles from 50 percent of the houses. It must be at least 3 miles from the factory. It must be within town limits.

Place a 1-1/2 square mile recreation area.

Question: Which area is greater—the area of town bordered by the town limit line and Green Street or the area of Clear River Park?

Test administrators presented these cards without giving instructions to students as to the order of the tasks or which equipment to use. The administrator answered questions about the goal of the activity but not about how the activity should be carried out.

SESSION 4C – MATHEMATICS OPEN-RESPONSE QUESTIONS

YOU MAY USE A CALCULATOR ON THIS SECTION.

For each of the following six questions, show all work as well as your answers in the spaces provided for these questions in your answer booklet. Show all diagrams, tables, computations, etc. that you use. If you do the work in your head, explain in writing how you did the work. CIRCLE your final answer to each question. DO NOT WRITE ON THIS PAGE!

TURN TO PAGE 8 IN YOUR ANSWER BOOKLET AND RECORD ANSWERS FOR QUESTIONS 1-5.

1. The diagram in your answer booklet is a scale drawing of John's room. John has four pieces of furniture that he needs to put in the room. The measurements of the furniture are:

Bed	6 feet long	and	3 feet wide
Desk	5 feet long	and	3 feet wide
Chest	5 feet long	and	2 feet wide
Bookcase	4 feet long	and	1 foot wide

When arranging the furniture John must follow these rules:

- The doors may not be blocked.
- Each piece of furniture must have at least one side against a wall of the room.
- Because the chest is too tall, it cannot be placed against a window.

The bookcase has been placed and labeled on the diagram. Choose a way that John could arrange the other three pieces of furniture so that the arrangement follows all the rules. On the diagram, show that arrangement by drawing in each piece of furniture. Draw each one to scale, using the same scale used to make the diagram. Label each piece of furniture.

2.

TELEPHONE CALLING RATES

	Day Rate Mon-Fri, 8am-5pm		Evening Rate Mon-Fri, 5pm-11pm Sat-Sun, 8am-11pm		Night Rate All Days, 11pm-8am	
From Allenville To	First Minute	Each Additional Minute	First Minute	Each Additional Minute	First Minute	Each Additional Minute
Burneyford	$ .09	$ .03	$ .07	$ .02	$ .05	$ .02
Camptown	$ .28	$ .09	$ .22	$ .07	$ .17	$ .05
Dorning	$ .37	$ .11	$ .30	$ .09	$ .22	$ .07
Edgeton	$ .42	$ .12	$ .34	$ .10	$ .25	$ .07

Ken made two telephone calls Monday from his home in Allenville. At 11:15 a.m, he called Tom in Burneyford and talked 15 minutes. That night at 6 p.m. he called Al in Dorning and talked for 5 minutes.

A. Explain how Ken will determine which call is more expensive.

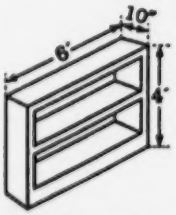
B. How much would Ken have saved by placing both calls between 11 p.m. and 8 a.m.?

3. Look at the number sentences below and describe the pattern.

53 x 111 = 5883
26 x 111 = 2886
43 x 111 = 4773
12 x 111 = 1332

4. At 10:03 a.m., you enter a parking garage. The parking rates are posted on the sign at the entrance. You leave the garage at 3:48 in the afternoon. If you give the parking attendant a ten dollar bill, how much change should you receive?

PARKING RATES
$ .75 first hour
$ .50 each additional hour
$5.00 daily maximum



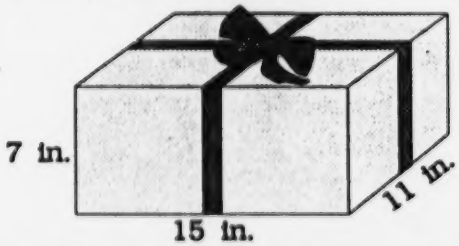
SALE PRICES
#2 Ponderosa Kiln-Dried Pine 1"x10"

LENGTH	8'	10'	12'
PRICE	$5.25	$6.30	$7.30

5. You plan to build the bookcase sketched above. The newspaper has an ad from a store which carries the 1" x 10" Ponderosa pine boards you want for the project. Use the ad to estimate what the lumber will cost. Explain the procedure you used and your reasoning.

TURN TO PAGE 6 IN YOUR ANSWER BOOKLET AND RECORD YOUR ANSWER FOR QUESTION 6.

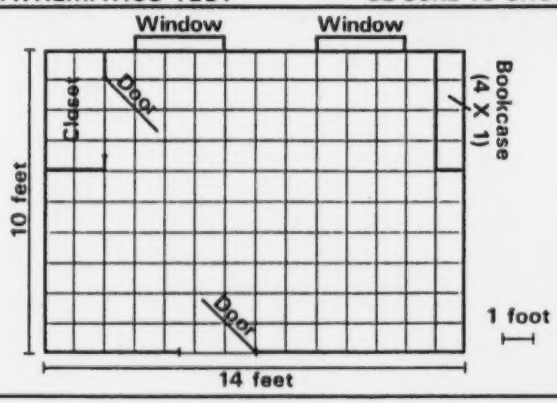
6. Susan has a package to wrap that has the dimensions shown below. The knot and bow require 14 inches of ribbon and the package is tied with ribbon all the way around as shown.



How much ribbon is needed? Please explain your reasoning.

PLEASE STOP! DO NOT GO ON TO NEXT PAGE.

1



Attachment 3b
MEA Open Ended

2

3

4

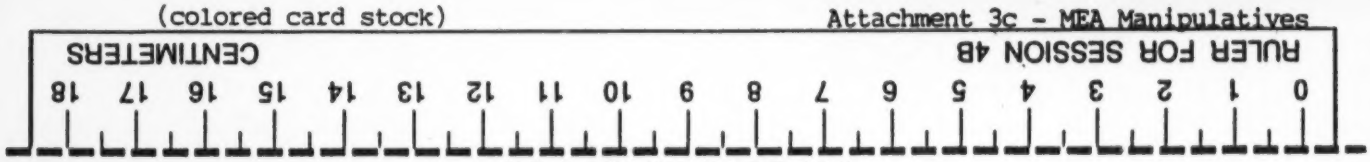
5

RECORD YOUR ANSWER FOR #6 ON PAGE 6 (SESSION 4C)

PLEASE DO NOT WRITE IN THIS AREA

1 ⓪①②③④○ 2 ⓪①②③④○ 3 ⓪①②③④○ 4 ⓪①②③④○ 5 ⓪①②③④○

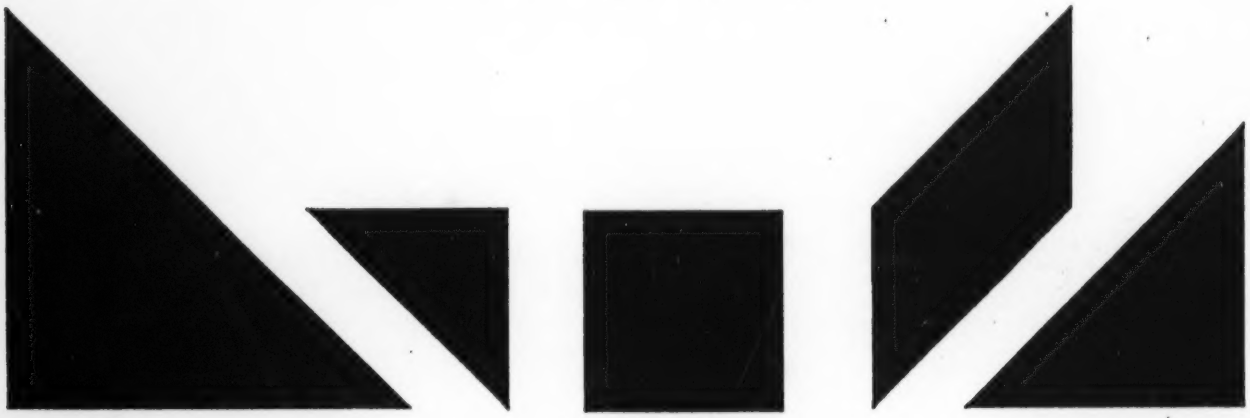
1051732



CUT ALONG THE DOTTED LINE

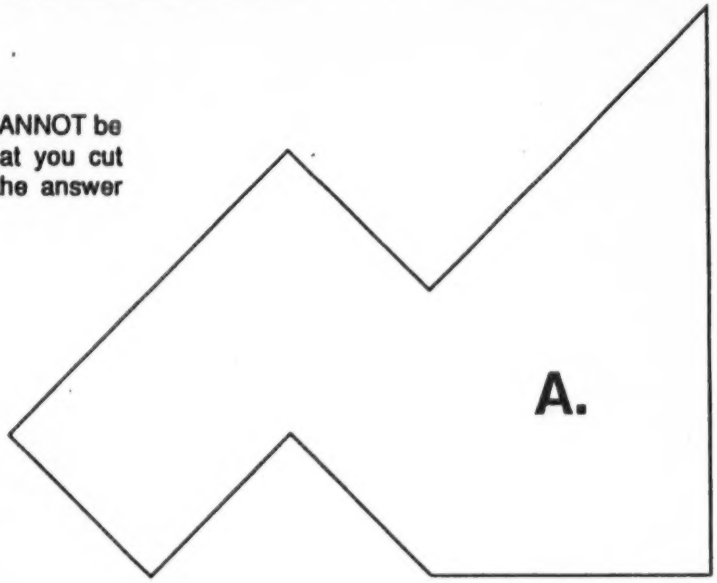
TANGRAM PIECES FOR QUESTION #41

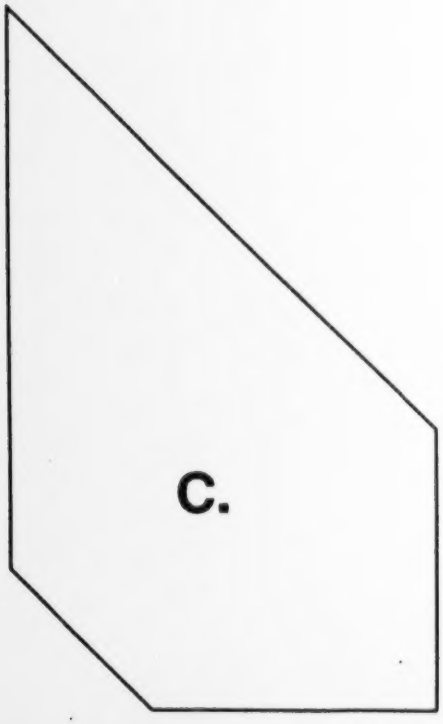
(Cut out each piece separately.)

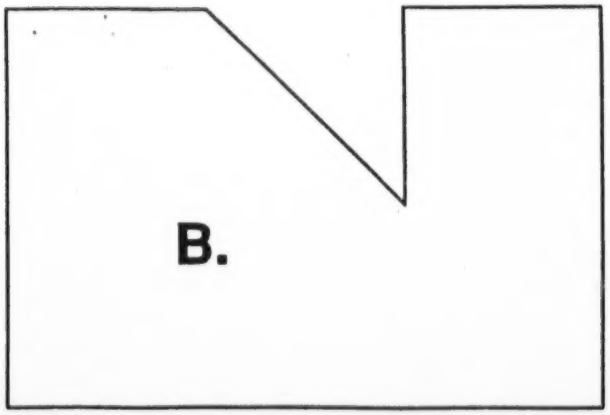


CUT ALONG THE DOTTED LINE

Question 41: Which of the following figures CANNOT be made using all five shapes that you cut out? (Record your answer in the answer booklet.)





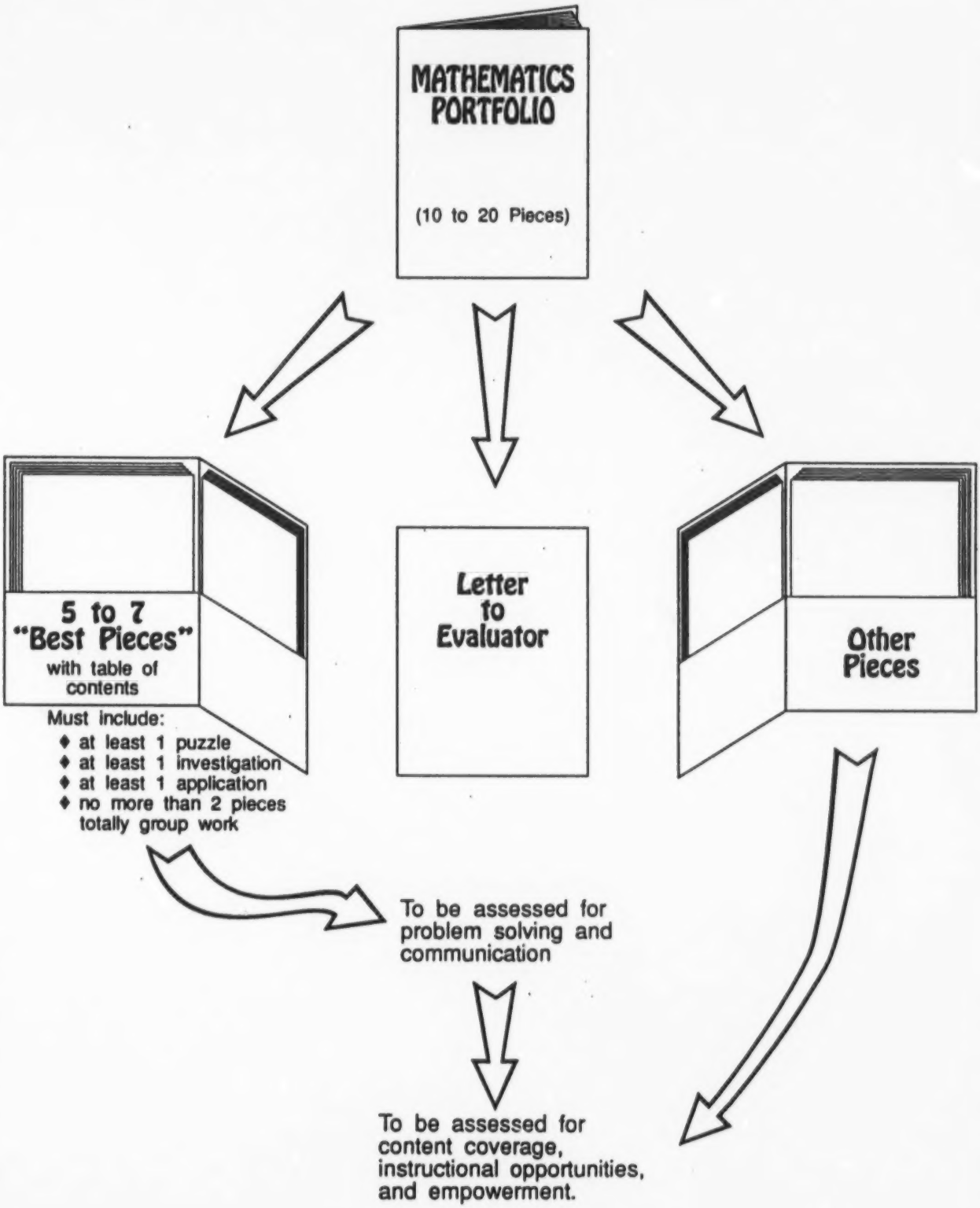


Attachment 3d - MEA Counters

(perforated card stock)

BEST COPY AVAILABLE

VERMONT MATHEMATICS PORTFOLIO: CONTENTS



VERMONT MATHEMATICS PORTFOLIO: ASSESSMENT

Rated by Criteria

Problem Solving

PS1 How well does the student understand the problem?

PS2 How does the student solve the problem?

PS3 Why does the student solve it the way s/he did?

PS4 What observations, connections, generalizations does the student make about the problem?

Communication

C1 What terminology, notation, symbols does the student use to communicate his/her math thinking?

C2 What representations (graphs, charts, tables, models, diagrams, pictures, manipulatives) does the student use?

C3 How clear is the student's communication of mathematical thinking and problem solving?

Assessed by Anecdotal Comments and Tallies

Content Areas

- Number Sense – Whole/Fractions
- Number Relationships/Theory
- Operations/Place Value
- Estimation
- Patterns/Functions/Relationships
- Algebra
- Geometry/Spatial Sense
- Measurement
- Statistics/Probability
- Logic

Empowerment

- Curiosity
- Flexibility
- Risk-Taking
- Perseverance
- Reflection

Instructional Opportunities

Evidence of Inclusion of:

- Group Work
- Interdisciplinary Work
- Manipulatives
- Real World Applications
- Technology

In summary, the Vermont Assessment Program includes two distinct snapshots that contribute to the overall picture of mathematics education in the State. **Best Pieces** within portfolios of individual students are used to assess the problem solving abilities and communication skills of students. **Portfolios** of student work provide a picture of the instructional opportunities, the content areas of programs, and anecdotal indicators of disposition. A detailed description of each of the components and the criteria that comprise the assessment are provided in this guide.

VERMONT MATHEMATICS PORTFOLIO PROFILE WORKSHEET

Student: _____

ID Number: _____

School: _____

Grade: _____ Date: _____

Rater: _____

	PS1 Understanding of Task	PS2 How – Approaches/Procedures	PS3 Why – Decisions Along the Way	PS4 What – Outcomes of Activities	C1 Language of Mathematics	C2 Mathematical Representations	C3 Presentation
	SOURCES OF EVIDENCE • Explanation of task • Reasonableness of approach • Appropriateness of response leading to inferences of understanding • Identification of assumptions	SOURCES OF EVIDENCE • Demonstrations • Descriptions (oral or written) • Drafts, scratch work, etc.	SOURCES OF EVIDENCE • Changes in approach • Explanations (oral or written) • Validation of final solution • Demonstration	SOURCES OF EVIDENCE • Solutions • Extensions – observations, connections, applications, synthesis, generalizations, extensions	SOURCES OF EVIDENCE • Terminology • Notations/symbols • Equations	SOURCES OF EVIDENCE • Graphs, tables, charts • Models • Diagrams • Manipulatives • Equations linked to other representations	SOURCES OF EVIDENCE • Audio/video tapes (or transcripts) • Written work • Teacher interviews/observations • Journal entries • Student comments on cover sheet • Student self-assessment
ENTRY 1 Title: _____ P Puzzle I Investigation A Application O Other							
ENTRY 2 Title: _____ P Puzzle I Investigation A Application O Other							
ENTRY 3 Title: _____ P Puzzle I Investigation A Application O Other							
ENTRY 4 Title: _____ P Puzzle I Investigation A Application O Other							
ENTRY 5 Title: _____ P Puzzle I Investigation A Application O Other							
ENTRY 6 Title: _____ P Puzzle I Investigation A Application O Other							
ENTRY 7 Title: _____ P Puzzle I Investigation A Application O Other							
OVERALL RATINGS →	UNDERSTANDING OF TASK FINAL RATING 1 Totally misunderstood 2 Partially understood 3 Understood 4 Generalized, applied, extended	HOW – APPROACHES/ PROCEDURES FINAL RATING 1 Inappropriate or unworkable approach/procedure 2 Appropriate approach/procedure some of the time 3 Workable approach/procedure 4 Efficient or sophisticated approach procedure	WHY – DECISIONS ALONG THE WAY FINAL RATING 1 No evidence of reasoned decision-making 2 Reasoned decision-making possible 3 Reasoned decisions/judgments offered with certainty 4 Reasoned decisions/judgments demonstrated	WHY – OUTCOMES OF ACTIVITIES FINAL RATING 1 Solution without extensions 2 Solution with observations 3 Solution with connections or applications 4 Solution with synthesis, generalizations, or abstraction	LANGUAGE OF MATHEMATICS FINAL RATING 1 No or inappropriate use of mathematical language 2 Appropriate use of mathematical language some of the time 3 Appropriate use of mathematical language most of the time 4 Use of rich, precise, elegant, appropriate mathematical language	MATHEMATICAL REPRESENTATIONS FINAL RATING 1 No use of mathematical representation(s) 2 Use of mathematical representation(s) 3 Accurate and appropriate use of mathematical representation(s) 4 Perceptive use of mathematical representation(s)	PRESENTATION FINAL RATING 1 Unclear (e.g., disorganized, incomplete, lacking detail) 2 Some clear parts 3 Mostly clear 4 Clear (e.g., well organized, complete, detailed)

COMMENTS:

EMPOWERMENT COMMENTS (risk taking, curiosity, flexibility, reflection, perseverance):

CONTENT AREAS

Number Sense – Whole No./Fractions, Number Relationships, Theory (N)

Operations/Place Value (+)
Operations (R)

Estimation (+R)

Patterns/Relationships (+)
Patterns/Functions (R)

Algebra (R)

Geometry/Spatial Sense (+R)

Measurement (+R)

Statistics/Probability (+R)

Logic (+R)

INSTRUCTIONAL OPPORTUNITIES

Attachment 4c

BEST COPY AVAILABLE

PERFORMANCE EVENT FACILITATOR INFORMATION SHEET

TASK: **M2 - SOUP CANS** **GRADE 12**

NUMBER IN GROUP: 3 - 4 students

OVERVIEW:

As a group, the students will brainstorm ideas for designing a soup can. The students will then work individually to design a new soup can according to the requirements given to them in their response forms. Each student is then encouraged to promote his or her can as the best choice using any methods available.

SET-UP MATERIALS:

- Place these materials in the middle of the table:
  - rulers
  - scissors, left and right
  - pencils
  - construction paper
  - scrap paper
  - compasses
  - poster paper
  - calculators
  - markers
  - tape
  - student response forms

OTHER INFORMATION:

Allow 15-20 minutes for the brainstorming. Students will be recording information in their response forms during this time. After the brainstorming session, direct the students to open their forms and complete the individual tasks. Each student in the group should have a response form which details a different type of can which they are to design. This work is to be done on an individual basis. Separate the students, if possible. (Note: There are 4 different versions of the second page of the student response form.)

PERFORMANCE EVENT FACILITATOR INFORMATION SHEET

TASK: **M6 - PEP CLUB FUND RAISER** **GRADE 12**

NUMBER IN GROUP: 3 - 4 students

OVERVIEW:

Data is given concerning a fund raising event. The students work as a group to determine several different methods of reporting the total sales and the number of awards to the club membership. Remaining in a group, each student will create a unique method of presentation. The final 10 minutes of the test period will be spent answering an individual question about whether the activity should be repeated again next year.

SET-UP MATERIALS:

- Place these materials in the middle of the table:
  - poster paper
  - colored pens or markers
  - rulers
  - yardsticks
  - scissors
  - compasses
  - protractors
  - clip art
  - pencils
  - scrap paper
  - tape
  - calculators
  - student response forms

OTHER INFORMATION:

This group will need a table or some large, flat area to spread out their materials.
10 minutes before the end of the test period, the facilitator will direct the students to complete the back (page 4) of their forms individually.